

(When Would We Call a Machine Stupid?)

When Will an Intelligent Mechanism Not Learn?

Surprise

CRISIS NOTES

CYBERNETICS

When certain types of events (e.g. an outbreak of anti-Americanism? SU sluggishness — or advance?)

come always as surprises, one can infer:

1) that the ~~abnormally~~ atypically low prior probs assigned to these are not changing with experience; no learning.

2) that the thresholds for reporting or responding to evidence concerning them must be high; bias, costs of errors, payoffs.

3) the model for interpreting evidence may be wrong

4) important hypotheses may be ignored, or not yet imagined (so that "warning" evidence — "strong" evidence likely to be received well before the event $\S$ — would not produce change in opinions, and response).

Analyses: BNSP as info

Unsigned draft of BNSP as info

(Note desire of "troops" to anticipate McN's reactions — not first to oppose them.)

When their desires are known to be counter to his, Inaction may seem to them a more logical (as well as safe) response to certainty about his wishes than taking initiative.)

Note McN's intentions: 1) to be ambiguous

2) to take actions contributing

to ambiguity, without intending.

3) to conceal from himself the extent of this ambiguity.

+ intentions of troops to snatch an unequivocal meaning from his words/acts.

C + C

Preference map ranks all "possible" (thought-of) end-positions, consequences, outcomes, goals : in a space whose dimensions are "value-criteria" OR "objective parameters" (if only one criterion; or if separate criteria, & relationships, ~~are~~ "basis for ordering" — are implicit).

[Th²o, if a dimension — physical or "value" — is omitted, map may be quite unreliable; and "true" map unexplored.

b) In parts of space regarded as "unlikely," "unattainable," map is tentative, "unreliable"; likely to change on closer inspection, as new actions or changes in environment make these outcomes "possible," "relevant."]

COMMAND decision: 1) Given a broad class of actions and broad set of "possibly attainable," "relevant," outcomes, determines preference map in that region (first, specifies dimensions of "value space" — for that region).

2) Specifies a ^{sub-}region of outcomes as "the goal," and/or a sub-class of actions (a "policy," "strategy") to be ^{examined} considered.
[low probs: 0 ~ 1 w.r.t. responses] [Policy + reasons]

C+C

Control : then selects one of the actions within the "policy" on further analysis + info; perhaps sequentially.

If both "policy" and "reasons" are given, then: 1) New info (or, difference of opinion) may indicate defects in reasons, e.g. predictions, that may ^{lead Control} question "policy" (as well as guiding selection within policy). Likewise, ordering of values may be questioned.

2) If no "reasons".

If this "feedback" to Command is not desired, then Policy may be specified without "reasons" — predictions + preferences. But then: 1) Control may ignore; ~~or~~ (not convinced that it "agrees");

2) Control will choose within policy by guessing at "goal," or will follow its own goals.

3) Control will not be able to adapt to info relevant to goals within policy, or warn Command of need to change policy.

C+C

§ (2) If "goal" is given without policy, and if Command has not analysed/predicted relation of alternative policies to its goals, it is unable to recognize or correct choices by Control that run counter to these goals: Control can take over Command functions (goal-setting).

~~Ex~~ May follow sequence: a) Command specifies broad goal-region; b) Control (staff) generates broad alternative policies ^(+ rough outcomes) (classes of action-sequences); c) Command refines preference-map in region of policies and chooses a policy.

Staffwork in (b) may be part of Command staff, or may be done by part of Control staff.

Cybernetics:

C. explores & analyzes certain similarities between organisms, organizations, and certain kinds of machines: similarities in structure and in dynamic behavior.

Gotten bad name by labelling too soon a new "science", what is really an example of poetry: the exploitation of analogies, metaphors: the generation & initial testing of ~~the~~ hypotheses or conjecture that certain analogies are likely to be fruitful, lead to understanding of structure and to demonstration of basic isomorphisms. [All science starts with such "poetry."]

A triple metaphor is involved; and different investigators are more expert & more interested in one or another of the triad; some want ^{to invent} machines that behave more like men (thinking, problem-solving, pursuing goal-seeking); some want to understand men; some want to understand organizations, and make them more like men. (Some might want to modify man, or help him modify himself.) Each has questions, uncertainties; they make progress toward answers unevenly. Basic to

Cybernetics is the perception (conjecture) that these lists are related; that questions on one list are likely to suggest useful questions for another; that answers (partial, tentative) to a question on one ^{list} are likely to be relevant to questions on both of the other; that new tools for solving one set are likely to be relevant to the others.

(Direct simulation may be possible).

By exploiting this "fact", progress in the three areas can be shared and accelerated.

First step is to evolve a common language, or rules for Translating problems & results; & to share concepts, perceptions of forms & patterns.

[Note: for investigators interested in aspects of one area that are not similar to others, this can be a costly waste of time. Hence, important to determine what the "fruitful metaphors" are.]

What are the properties of machines that are relevant? (i.e. what defines the class of machines that are relevant?)

Original answer: a) Feedback ("

a b) negative feedback

a c) dynamically stable. ("teleological")

Now: receiving, transmitting, responding to information.

It is little noted that if "information" is defined operationally, this new def. is likely to identify a different class of machines from original: excluding those that can't be said to process info. internally (pendulum; ball in bowl) or to be "responding to info" in achieving equilibrium (cat in bath, dying?); including some that do respond to info but fail to achieve equilibrium, and some that rely more on "feedforward," indirect evidence, reasoning & prediction than on "feedback."

[We want machines/organizations that learn, and learn not only from "past results" (in contrast to "learning theory"). Conceivably, organism can learn faster & more precisely than it can adapt responses

Cyb. We want machines to serve our purposes
better by acting like "intelligent slave"; ^{gathering & utilizing info, working} without direction, ^{learning from} mistakes).

We want machines to learn; guess; form hypotheses;
test, utilize info from varied sources (including
feedback); and form & modify sub-goals, values;
organize itself, split up work among sub-units,
plan, choose sequence of actions, tests, observations;
recognize, infer.

- [Those interested in man may find metaphor useful:
- 1) because study of machines/orgs may suggest patterns of behavior as yet unnoticed in man;
 - 2) because similarity of behavior may indicate similarity of structural, process.

But: at this stage, it is still best to regard
the notions of "purpose, intent, expectations, goal,

as metaphors when applied to machines or orgs, whose
"validity" remains to be demonstrated, explored.

They have been borrowed on basis of the "fact" (conjecture)
that some machines exhibit ^{some of} the behavior which, if observed
in human, lead us to infer the existence & nature of his
purpose, expectations, values, goal.

Mocky "INTROSPECTIVE BEHAVIORIST"

(Question remains: are certain cues to our use of these terms for a human being overlooked (other than "humaneness")? so that usage is misleading.

Are patterns really similar, or are differences in the machine behavior being overlooked? To what extent do similar structures, processes, underly the similarities in behavior? Is the metaphor useful? Dangerous?

One thing is clear: ^(like) if a human is a machine,
he is not a ^(like) simple machine.

i.e. comparisons to "simple" machines — even, to simple "goal-seeking" machines (thermostat; pendulum?) — are unlikely to be very helpful to the study of either (except in the gross way in which study of anatomy, or economy, can be likened to astronomy or geography). Study of differences likely to be more fruitful than similarities, but even this, not very helpful.

[Whereas, in proper sphere of comparison, study of differences is likely to be helpful: in part, to combat effects of unconscious metaphors.]

Emphasis on "probability" as frequency is like emphasis on feedback as main or sole source of relevant information: or rather "reflex/reinforcement" theory

[Adaptive controller influenced by memory uses more than "feedback": it must "conjecture" that given past results are relevant, similar enough, to current situation — if it is to avoid "stupidity," "stereotyping." And, info. w. internal structure is not "feedback"; needed, unless internal adaptation is to be made "blindly," "randomly."]

↓ takes perspective of repetitive behavior, sequence of actions, where only average or "final" (equil.) actions/results matter (perhaps because commitments can be deliberately limited — risks limited — during early sequence of trial moves).

Attitude: "Try anything once; Why Not?; Let's see what happens; ... That's how you learn; School of Hard Knocks; Not to learn from mistakes; You can't win them all." (Formulas to make actor willing to risk error; reduce his anticipation of regret; minimize cost of error, emphasize benefit, in info, learn of error — improvement in later expectations).

This simply is not appropriate framework or perspective in which to view decision-situations in which :

- 1) Major risks involved in commitment;

- 2) Little opportunity for posthoming parts of decision, making it sequential; or

- 3) Little hope of info coming in prior to major "outcome" (in time to adapt to);

- 4) Single mistake can be disastrous; at least, drastically change the problem, for the worse; irrevocable; perhaps, disintegration of org, problem-solving capability; perhaps, lead to change in decision-making personnel (+ perhaps, goals, images).

Can't wait for trial "results" to make choice among commitments; or, can't afford costly "trial", or afford risks of trial (until non-trial info available).

Here: must utilize a variety of info, from various kinds of one's + others' past experience, borrow hypotheses, look for "clues" as to future "results" from ~~the~~ current info other than info on own "results" of one's current or recent past behavior.

Same applies when "feedback" is slow, lagged (compared to changes in system - environment) or unreliable (noisy channel) or equivocal (inadequate theory relating behavior to observations on behavior).

Over-reliance on "feedback", compared to other info, may be simply misleading or dangerous.

(esp. if it leads to failure to collect or utilize other info). Or much more costly, slow, inaccurate.

Open-loop control in Atomic Control in fact reflects recognition of these possibilities.

On the other hand, perception of situation in above terms may also be in error, or exaggerated. (R&D; control). If advantages of feedback were recognized (and actual inadequacies of other sources of info, or theories), Research might reveal that informative actions, tests, trials, experiments, with limited risk, can be taken in time, without seriously compromising possibility of achieving acceptable results; decision can be made sequential, some aspects can be postponed (kept uncertain)

"Testing the water," ... "Let's see how it goes"

This may involve merely treating early stages of process as "experiment," and preparing to learn from them: to receive + interpret info, and adapt to it.

Even when decision is "single-shot," "big," recognition that there ^(may) will be other decisions may encourage and act on to prepare to learn from (even disastrous) outcome:

e.g. 1) keep certain records

2) know "what was done" (2 men defusing a new type of mine)

3) have resources for analysis; &

4) observe ^{if come} results in some detail; station observers

5) record times of events; to allow later examination of role of unrecorded, outside influences on result; & to discover "system response characteristics";

6)

All this has costs; may utilize resources that could be used in present "crisis."

Crisis

To suggest that you should invest resources to learn from ^{next} crisis is to propose that it act as if:

- a) there might be another crisis; and
- b) there might be another one after that.

(Expectation that the crisis-after-next will be ~~not~~ be "someone else's problem" reduces incentive to prepare now to learn from next crisis, to ~~its~~ improve performance in the crisis-after-next.

One can still try to learn from past crises — if one anticipates one more — but if "learning ~~from~~ mechanism" has never been instituted, past records are likely to be inadequate, equivocal, even misleading (e.g., records based only on public, unclassified sources; or, "Command" data but no operational or Control data).

(Similarly — Jones — Commander who anticipates that a postponed decision will throw interpretation of info + choice of action to someone else will be reluctant to make decision sequential.)

Cybernetics

We are not used to defining "purpose," "desire, intent, goal, opinion" behaviorally. First attempts to do so might lead to "misuse" of term. Still, we can ask:

"What observations of another person's behavior (and its "results" — including its observable results on us, e.g. on our emotions) lead us to infer that he is acting purposively, to achieve a certain goal?"

It is then possible that a machine could be constructed to exhibit all these behavioral characteristics.

Resistance to this identification can be fruitful, if it leads us to search for new, "essential" cues, criteria, (unless we short-circuit this by insisting on "humanness" or closely human ^{or animal} structure). We might find that behavioral cues were never enough, even with human.

E.g., one behavioral correlate of human "choice, deliberation, problem-solving" is that it tends to be unpredictable by observer. Machine can do that. Is it "enough"? $\gg$ (Is analogy of cognition to internal noise useful?)

We use similar metaphors when we use
use concepts derived from introspection, consciousness/
observation of self, to:

- a) other humans
- b) animals
- c) ~~xxx~~ plants
- d) organizations ; nations ; ^{e)} "peoples, races"
- e) natural processes.

f) organs of our body.

In some cases — e.g. C, f — we are more

conscious of metaphor, "anthropomorphism": the wind
caressing, sea lapping, grass sighing, plant seeking
(we don't expect analogy to be fruitful of new questions
or concepts). In others, analogy is almost irresistible
irresistible = other humans, animals (fear, love,
hunger).

But analogies to machines have been resistable;
partly because they are so slavish (Mackay), partly
man-made (without honor in the home town), but
mainly because they have not been "very similar".

[Marsden: Orig. "chapter" as Mackay's CPM.]

In fact, the "new machines" that do show behavioral resemblances to man are either recent or not constructed yet.

(Nor are they really likely to be unless we can exploit the analogy; our everything we can find out about man is the design, rather than having them imitate observable behavior of rats or neurons, or letting them "grow" from present machines: Minsky).

[Note that "reverse anthropomorphism" is not entirely new; many sophisticated concepts with which we interpret our own behavior are derived from observations of non-human systems: groups ("conflict"), instruments, hydraulic systems (blood; liquids)

Note Bartlett on role of "metaphor".

COMM

Analyze: constraints ^{design, reporting of} experiment whose results
are to be used by someone else.
(Science).

(e.g., myself later; or, new Administration;
or, other parts of Administration; or Ally;
enemy.)

"Demonstrations" as Messages

as Experiments (results presumed
highly predictable by experimenters;
needed to convince, demonstrate.)

Goals & sub-goals

A measure of agreement on social policy and agreement on procedures is both necessary & sufficient for coordinated action;

agreement on goals or predictions is not.

(It may be that some expectation of agreement on policy — based on perception of agreement on goals, opinion — is necessary to motivate search for common policy; likewise, to promote identification, desire to help (not merely to use, or cooperate with).

Fair Division problem.

But, pursuit of my goal must be at least compatible (given available policies) with maintenance of your expectation of achieving one of your goals (e.g. receiving payment, avoiding punishment).

Contrast: lack of cooperation of Ukrainians with Nazis. (Or, people who try to "use" Communists).

"Feedback learning," like "reinforcement learning," inadequate, too slow, inflexible, or unlikely to survive, if:

- 1) environment too complex

cues too equivocal, or too

- 2) goals complex, interdependent

- 3) no contours in environment, relevant to criterion ("altimeter"). "Results" not labeled (AN) ~~the~~ "reward" or "getting warmer."

- 4) environment hostile or dangerous; also, reaction dangerous.

- 5) Unavoidable lags in feedback, too long relative to tempo of change of environment, + range of ^{correction} responses available.

In some of these circumstances, nothing will work very well, or ensure safety or success; but probably helpful or essential to make trial-behavior depend on memory of cases before last one, and "analogies," and "other info."

(though some randomness — apprehensibility even to actors, who cannot supply reasons — may still be useful).

Unless calculation is adequate to bring actual output into neighborhood of goal, so feedback can be used for fine adjustment (control), feedback will show merely the existence of "gross error."

Some costs/risks of making decision sequential,

- { 1) Delay
- { 2) Lower reward

but not necessarily relative to actions — commitments — actually contemplated at outset. Info-gathering may show outcome better or faster than any of earlier commitments, or show that none of them would have succeeded.

When detective says, "One of you is the killer," the search process is almost over. Yet this is where Bayesian model approach begins (with adequate model).

To try this game — hoping for "conclusive" breakdown under tension, giveaway — or "suspects" drawn at random would be slow & unpromising.

(Real detectives ~~that~~ can't afford to start with the better).

Feedback:

~~To call~~ feedback can be called "negative" unambiguously only when response variable is one-dimensional (fuel supply, heat, power input...): goes only "up or down," "on-off," "back-forth," thermostat.

and "sign" of ~~correct~~ "correct" (error-reducing) response is always opposite to the sign of the error. (a to, rate of change of error).

But if response-variable is a vector, then responses will not have a definite sign (except: partial ordering); and more than one response may be "correct,"

[Does dimensionality of error matter?]

[Isn't it necessary to attach sign to ^{change in} error?

To be able to say it is increasing or decreasing — i.e. ordering? yes.

How about: is it necessary to attach sign to error, to say that it is positive or negative?

To assign metrics?

None of these are possible in early stages of hand problems: all that is fed back is existence of error.

Analyze: "trial and error" as a method.

Dependence

Out of a set of possible sample results, the Bayesian ordering of results in terms of "strength of evidential support for hypothesis A" depends on the other hypotheses considered.

Hyp. A: $p(\text{heads}) = .7$ 10 tosses

Strongest result is 7 heads, if alternatives are $.1, .2, \dots, .9$

But strongest result is 10 heads if alternative is only $p = .3$.

[Might it depend on sample space?

e.g. Does use of Significance Test lead to violation of Independence of Irrelevant Alternatives, in our "choices" or "preferences" among sample results in terms of support of given hypothesis?!

[But — might this be OK in context of Group Decision, as in statistics?!] [Hence, might it be OK for "vague" individual?]

This does not hold for Bayesian ordering of strength of results for Hyp. A vs. Hyp. B.

Inference:

Almost no "hypotheses" outside statistics are "simple stat. hyps" or composites of simple stat. hyps: hyps that imply (in fact, are defined by) precise likelihoods, with consensus (conventional, traditional, "objective")

Hence, one can speak of "learning" or modifying or making precise, the likelihoods associated with a given hyp.

Only statisticians are interested in "simple stat. hyps," i.e. in "choosing among propositions that (nearly) assert conditional probs"

[except, Intelligence Analysts?]

(estimates)
Such statements are rare examples of actions that are sensitive to precise probs.

Thesis = (see Lindblom, p. 40)
March & Simon, 137

1. Applies only to special conditions of uncertainty:
known sets of consequences (relevant events, hypotheses)

2. Vagueness & Terminal Decision; no experiments,
^{terminal} no more info [Applies to design of experiments]

3. Alternatives given;
Utilities given (goals, values)

4. Single individual. (^{vagueness gives} But: model of group decision)

(But: applies to nearly all cases to which Bayesian approach applies)

B&L: 45 there are "conditions whose attainment would imply that a solution was at hand."

see: 50

McN

Revisited:

BNSP + SecDef memos

We like "policy" being formulated in context, & on occasion, of concrete decisions.

P.J. SecDef memo to Pres. Then: need a SecDef memo ^{including CW} on Plans/use policy, not just on Budget. (RVD?)

But: should such a policy statement pretend to be comprehensive? Or, merely to be adequate to determine the specific issues raised by the current alternatives?

(to explain/justify these choices; and to be applied to future choices "sufficiently comparable," under "sufficiently comparable circumstances.")

(Thus, admit/specify the range of alternatives actually at issue, and the critical points of comparison (e.g., proposals by Services) rather than pretend that rules governing these choices were formulated independently of these choices, & the choices merely deduced from these rules.

Thus: don't have a BNSP (unless, in broad terms, and to in context of specific plans/use issues); but don't aim or claim comprehensiveness in SecDef memos.

Don't say "Characteristic A" (damage-limiting) is the ^{P.J.} general criterion of all choices in this area (e.g., force size).

Inference

Decision Theory and Value of Info approach considers experiments + info (only) in context of DT: given actions, values, hyps, etc.

Actually, goal, effect, + value of exp. + info may be in broader, prior context....

To imagine ~~that~~ these things as "hypothetically" given is to distract all attention from fact that they are not actually given.)

DT is great compared to no theory; but...

↳ in generating new insights (some of which should be regarded skeptically).